Why Occam's Razor

Russell K. Standish
High Performance Computing Support Unit
University of New South Wales
Sydney, 2052
Australia
R.Standish@unsw.edu.au
http://parallel.hpc.unsw.edu.au/rks

Abstract

In this paper, I show why in an ensemble theory of the universe, we should be inhabiting one of the elements of that ensemble with least information content that satisfies the anthropic principle. This explains the effectiveness of aesthetic principles such as Occam's razor in predicting usefulness of scientific theories. I also show, with a couple of reasonable assumptions about the phenomenon of consciousness, that quantum mechanics is the most general *linear* theory satisfying the anthropic principle.

03.65.Bz,01.70.+w

Typeset using REVTEX

I. INTRODUCTION

Wigner [1] once remarked on "the unreasonable effectiveness of mathematics", encapsulating in one phrase the mystery of why the scientific enterprise is so successful. There is an aesthetic principle at large, whereby scientific theories are chosen according to their beauty, or simplicity. These then must be tested by experiment — the surprising thing is that the aesthetic quality of a theory is often a good predictor of that theory's explanatory and predictive power. This situation is summed up by William de Ockham "Entities should not be multiplied unnecessarily" known as Ockam's Razor.

We start our search into an explanation of this mystery with the *anthropic principle* [2]. This is normally cast into either a weak form (that physical reality must be consistent with our existence as conscious, self-aware entities) or a strong form (that physical reality is the way it is *because* of our existence as conscious, self-aware entities). The anthropic principle is remarkable in that it generates significant constraints on the form of the universe [2,3]. The two main explanations for this are the *Divine Creator explanation* (the universe was created deliberately by God to have properties sufficient to support intelligent life), or the *Ensemble explanation* [3] (that there is a set, or ensemble, of different universes, differing in details such as physical parameters, constants and even laws, however, we are only aware of such universes that are consistent with our existence). In the Ensemble explanation, the strong and weak formulations of the anthropic principle are equivalent.

Tegmark introduces an ensemble theory based on the idea that every self-consistent mathematical structure be accorded the ontological status of *physical existence*. He then goes on to categorize mathematical structures that have been discovered thus far (by humans), and argues that this set should be largely universal, in that all self-aware entities should be able to uncover at least the most basic of these mathematical structures, and that it is unlikely we have overlooked any equally basic mathematical structures.

An alternative ensemble approach is that of Schmidhuber's [4] — the "Great Programmer". This states that all possible programs of a universal turing machine have physical existence. Some of these programs will contain self-aware substructures — these are the programs deemed interesting by the anthropic principle. Note that there is no need for the UTM to actually exist, nor is there any need to specify which UTM is to be used — a program that is meaningful on UTM_1 can be executed on UTM_2 by prepending it with another program that describes UTM_1 in terms of UTM_2's instructions, then executing the individual program. Since the set of all programs (infinite length bitstrings) is isomorphic to the set of whole numbers $\mathbb{N}$, an enumeration of $\mathbb{N}$ is sufficient to generate the ensemble that contains our universe. The information content of this complete set is precisely zero, as no bits are specified. This has been called the "zero information principle".

In this paper, we adopt the Schmidhuber ensemble as containing all possible descriptions of all possible universes, whilst remaining agnostic on the issue of whether this is all there is.[1] Each self-consistent mathematical structure (member of the Tegmark ensemble) is completely described by a finite set of symbols, and a countable set of axioms encoded in

[1] For example, this ensemble does not include uncomputable numbers — but should these peculiar mathematical beasts be accorded physical existence?

those symbols, and a set of rules (logic) describing how one mathematical statement may be converted into another.[2] These axioms may be encoded as a bitstring, and the rules encoded as a program of a UTM that enumerates all possible theorems derived from the axioms, so each member of the Tegmark ensemble may be mapped onto a Schmidhuber one.[3]. The Tegmark ensemble must be contained within the Schmidhuber one.

An alternative connection between the two ensembles is that the Schmidhuber ensemble is a self-consistent mathematical structure, and is therefore an element of the Tegmark one. However, all this implies is that one element of the ensemble may in fact generate the complete ensemble again, a point made by Schmidhuber in that the "Great Programmer" exists many times, over and over in a recursive manner within his ensemble. This is now clearly true also of the Tegmark ensemble.

II. UNIVERSAL PRIOR

The natural measure induced on the ensemble of bitstrings is the uniform one, i.e. no bitstring is favoured over any other. This leads to a problem in that longer strings are far more numerous than shorter strings, so we would conclude that we should expect to see an infinitely complex universe.

However, we should recognise that under a UTM, some strings encode for identical programs as other strings, so one should equivalence class the strings. In particular, finite strings (ones in which the bits after some bit number n are "don't care" bits) are in fact equivalence classes of all infinite length strings that share the first n bits in common. These strings correspond to halting programs under the UTM. One can see that the size of the equivalence class drops off exponentially with the amount of information encoded by the string. Under a UTM, the amount of information is not necessarily equal to the length of the string, as some of the bits may be redundant. The sum

$$P_U(s) = \sum_{p:U \text{ computes } s \text{ from } p \text{ and halts}} 2^{|p|}, \tag{1}$$

where $|p|$ means the length of p, gives the size of the equivalence class of all halting programs generating the same output s under the UTM U. This measure distribution is known as a *universal prior*, or alternatively a Solomonoff-Levin distribution [5]. We assume the *self-sampling assumption* [6,7], essentially that we expect to find ourselves in one of the universes with greatest measure, subject to the constraints of the anthropic principle. This implies we should find ourselves in one of the simplest possible universes capable of supporting self-aware substructures (SASes). This is the origin of physical law — why we live in a mathematical,

---

[2]Strictly speaking, these systems are called recursively enumerable formal systems, and are only a subset of the totality of mathematics, however this seem in keeping with the spirit of Tegmark's suggestion

[3]In the case of an infinite number of axioms, the theorems must be enumerated using a dovetailer algorithm.

as opposed to a magical universe. This is why aesthetic principles, and Ockam's razor in particular are so successful at predicting good scientific theories. This might also be called the "minimum information principle".

There is the issue of what UTM U should be chosen. Schmidhuber sweeps this issue under the carpet stating that the universal priors differ only by a constant factor due to the compiler theorem, along the lines of

$$P_V(s) \geq P_{UV} P_U(s)$$

where P_{UV} is the universal prior of the compiler that interprets U's instruction set in terms of V's. The inequality is there because there are possibly native V-code programs that compute s as well. Inverting the symmetric relationship yields:

$$P_{UV} P_U(s) \leq P_V(s) \leq (P_{VU})^{-1} P_(U)(s)$$

The trouble with this argument, is that it allows for the possibility that:

$$P_V(s_1) \ll P_V(s_2), \text{ but } \mathrm{P_U(s_1) \gg P_U(s_2)}$$

So our expectation of whether we're in universe s_1 or s_2 depends on whether we chose V or U for the interpreting UTM.

There may well be some way of resolving this problem that leads to an absolute measure over all bitstrings. However, it turns out that an absolute measure is not required to explain features we observe. A SAS is an information processing entity, and may well be capable of universal computation (certainly *homo sapiens* seems capable of universal computation). Therefore, the only interpreter (UTM) that is relevant to the measure that determines which universe a SAS appears in is the SAS itself. We should expect to find ourselves in a universe with one of the simplest underlying structures, according to our own information processing abilities. This does not preclude the fact that other more complex universes (by our own perspective) may be the simplest such universe according to the self-aware inhabitants of that universe. This is the bootstrap principle writ large.

III. THE WHITE RABBIT PARADOX

A criticism levelled at ensemble theories is to consider universes indistinguishable from our own, except for the appearance of something that breaks the laws of physics temporarily, e.g. a white rabbit is observed to fly around the room at specific time and place.[4] There are two possible explanations for this:

1. that there is some previously unknown law of physics that caused this rather remarkable phenomenon to happen. However, it would have to be an extremely complex law, and thus belong to a rather unlikely universe.

[4] This problem was first discussed in Marchal [8], where it is called the *Universal Dovetailer Paradox*. Marchal used the term "White Rabbit" in [9], presumably in a literary reference to Lewis Carrol.

2. that there is some "glitch" or "bug" in the program governing the universe, that allows some of the "don't care" bits to be interpreted. Since there are many more ways a program can fail, than be correct, surely then, the "White Rabbit" universes should outnumber the lawlike ones.

Consider more carefully the latter scenario. In most of the universes where the "don't care" bits are interpreted, the "don't care" bits will be devoid of information, and appear as random noise to the self-aware entity, and thus the universe is indistinguisable from a law-like one. Only when the "don't care" bits form a pattern recognisable by the self-aware entity, will a breakdown of physical laws be observed (such as seeing a flying white rabbit). Such patterns, of course, will be sparse in the space of all such "don't care" bitstrings, and so the vast majority of the pathological universes would be indistinguishable from the law abiding universe they approximate.

Another viewpoint on this explanation is to realise that SASes are themselves finite entities, with finite discriminatory powers and memory. Therefore, the SAS imposes an interpretation filter on the data perceived from the universe it inhabits, imposing order (or compressibility) in its interpretation of the universe, even if no such order exists. Even though incompressible strings vastly outnumber compressible ones, the "interpretation filter" of the SAS maps these incompressible strings onto compressible ones. This implies a large number of "don't care bits" in any description of a universe, with correspondingly larger numbers of "don't care bits" for simpler descriptions, giving rise to the universal prior. Any white rabbit universe must therefore take on the appearance of being a consequence of complicated physical law, (i.e. case 1 above) which must be rare accoding to the universal prior.

Marchal's universal dovetailer paradox is expressed somewhat differently to the preceding description of the white rabbit paradox. He assumes that all SASes are neither more nor less than universal turing machines, and concious experiences are implemented as computations. This is a form of strong AI he calls *COMP*. The set of all possible computational continuations can be generated by the dovetailing algorithm. Since all such continuations exist, and bizarre experiences (eg the white rabbit) by far dominate the numbers of continuations, the paradox is why we experience order in the world.

This specification of the problem does not admit an obvious measure on which to decide which experiences are more likely than others. However, each such computational continuation can be identified with a string from the Schmidhuber ensemble, so the universal prior is defined over the set of such experiences, and the above arguments about the general white rabbit problem also apply to the universal dovetailer paradox.

IV. QUANTUM MECHANICS

In this section, I ask the question of what is the most general (i.e. minimum information content) description of an ensemble containing self-aware substructures. Firstly, it seems that time is critical for consciousness — i.e. in order for there to be a "flow of consciousness". Denote the state of an ensemble by ψ. This induces an evolution equation

$$\frac{\partial\psi}{\partial t} = \mathcal{H}(\psi, t) \tag{2}$$

Now conscious observers induce a partitioning for each observable $A : \psi \longrightarrow \{\psi_a, \mu_a\}$, where a indexes the allowable range of "classical" observable values corresponding to A, and μ_a is the measure associated with ψ_a ($\sum_a \mu_a = 1$[5]). The ψ_a will also, in turn, be solutions to equation (2).

If we further assume that the states ψ are elements of a vector space, and that the evolution equation (2) is linear, we may write $\psi = \sum_a \mu_a \psi_a$, and the observable operators can be written compactly as a linear operator $\sum_a \psi_a \psi_a^\dagger$, where $\psi\dagger$ is the dual of ψ. Measure is clearly given by $\mu_a = \frac{\psi_a^\dagger \psi}{\psi_a^\dagger \psi_a}$. Furthermore, since scaling does not change the physical state represented by ψ, we can assert without loss of generality that $\frac{\partial}{\partial t}(\psi^\dagger \psi) = 0$, implying that $\mathcal{H} = iH$, where H is Hermitian, and ψ is a vector in a Hilbert space. The most general Hilbert space is one over the field of complex numbers. In short, by means of 3 assumptions, 2 of which appear to be irreducible properties of consciousness, and the third being that of linearity, Quantum Mechanics is derived from the anthropic principle applied to an ensemble.

Returning then, to the issue of linearity. This is not an obvious requirement for anthropic universes, so must have an explanation. Weinberg [10,11] experimented with a possible non-linear generalisation of quantum mechanics, however found great difficulty in producing a theory that satisfied causality. This is probably due to the nonlinear terms mixing up the partitioning $\{\psi_a, \mu_a\}$ over time. It is usually supposed that causality [3], at least to a certain level of approximation, is a requirment for a self-aware substructure to exist.

V. ACKNOWLEDGEMENTS

I would like to thank the following people from the "Everything" email discussion list for many varied and illuminating discussions on this and related topics: Wei Dai, Hal Finney, Gilles Henri, James Higgo, George Levy, Alastair Malcolm, Christopher Maloney, Jaques Mallah, Bruno Marchal and Jürgen Schmidhuber.

[5] Here, as elsewhere, we use Σ to denote sum or integral respectively as a is discrete of continuous

REFERENCES

[1] E. P. Wigner, *Symmetries and Reflections* (MIT Press, Cambridge, 1967).

[2] J. D. Barrow and F. J. Tipler, *The Anthropic Cosmological Principle* (Clarendon, Oxford, 1986).

[3] M. Tegmark, Annals of Physics **270**, 1 (1998).

[4] J. Schmidhuber, in *Foundations of Computer Science: Potential-Theory-Cognition*, Vol. 1337 of *Lecture Notes in Computer Science*, edited by C. Freska, M. Jantzen, and R. Valk (Springer, Berlin, 1997), pp. 201–208.

[5] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, 2nd ed. (Springer, New York, 1997).

[6] J. Leslie, *The End of the World* (Routledge, London, 1996).

[7] B. Carter, Phil. Trans. Roy. Soc. Lond. **A310**, 347 (1983).

[8] B. Marchal, in *Proceedings of WOCFAI '91*, edited by M. de Glas and D. Gabbay (Angkor, Paris, 1991), pp. 335–345.

[9] B. Marchal, Technical Report No. TR/IRIDIA/95, Brussels University (unpublished).

[10] S. Weinberg, Annals of Physics **194**, 336 (1989).

[11] S. Weinberg, *Dreams of a Final Theory* (Pantheon, New York, 1992).